Stata爬虫分享

Yi Wang

Sep11, 2021

什么是爬虫

- 爬虫就是自动抓取网页信息的代码。
- 爬虫工具：R，Python和Stata：
  - Python和R爬虫功能已较为成熟，拥有一系列实现各类功能的Library和package；
  - Python：Requests、lxml、BeautifulSoup、Scrapy等；
  - R：rvest包在爬取静态网站数据方面功能很强大，还有RCurl和XML包等。
- 相比而言，Stata虽然在数据处理和计量分析方面非常高效，但在爬虫方面可能还处于“爬行”阶段。

为什么要用Stata做爬虫

- Stata爬虫程序简单，有Stata编程基础可以很快上手，而且相同的程序可以很方便的在不同计算机上运行。
- 直接将抓取的数据保存为Stata格式，方便在做实证时进行调用。
- Stata14以来，字符处理功能已大幅改善，再配合copy和curl等命令，Stata的爬虫能力日渐强大。

Stata爬虫的基本步骤

- 网页分析：提取链接，使用copy或curl获取网页源代码；
- 请求并读入：将把含有所需数据的源代码无乱码读入Stata；
- 处理数据：主要是对字符串进行处理，提取源代码中的信息；
- 多网页的抓取。

Stata爬虫的必备技能

- 文本文件读入：infix等；
- 乱码处理：unicode命令、ustrfrom()函数等；
- 文本信息的处理：字符串函数、正则表达式；
- 宏与循环；
- HTML基础知识及解析网页；
- Python。

Stata爬虫案例

- 新浪财经上市公司高管任职数据爬取
  - 某上市公司页面

Stata爬虫案例

- 新浪财经上市公司高管任职数据爬取
  - 该上市公司某高管简历页面

长江电力 上海 600900

20.04 -0.29 -1.43% 2021-09-09 10:00:00

昨收盘:20.33 今开盘:20.50 最高价:20.50 最低价:19.96 代码/名称/拼音 查询 代码检索

市值:4557.47亿元 流通:4557.47 成交:168010手 换手:0.07% 公司资料意见反馈

公司资料： 公司简介 股本结构 主要股东 流通股股东 基金持股 公司高管 公司章程 相关资料

个人简介

姓 名	性 别	出生日期	学 历	国 籍
陈国庆	男	196410	博士	中国
简 历	陈国庆，男，1964年10月出生，总经理、党委副书记，工学博士，教授级高级工程师。历任葛洲坝电厂大江分厂电修车间计算机班班长、自动班班长，葛洲坝电厂大江分厂维修车间专责、副主任、党支部书记、主任，葛洲坝电厂大江分厂副厂长，中国长江电力股份有限公司三峡电厂副总工程师兼生技部主任、副厂长、党委委员，中国长江电力股份有限公司总经理助理兼副总工程师、总工程师、党委委员、副总经理，中国长江电力股份有限公司党委书记、副总经理。现任中国长江电力股份有限公司董事、总经理、党委副书记，兼任中国长江电力股份有限公司长电学院院长。		

The end, Thank you!